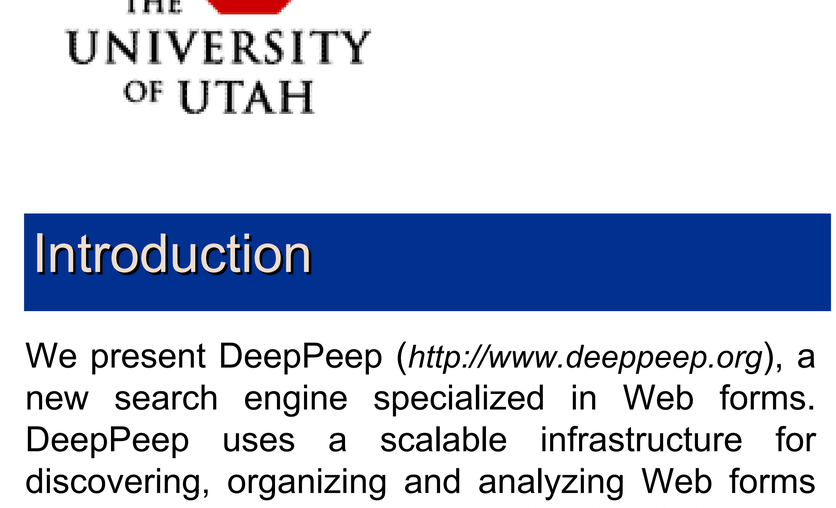
DeepPeep: A Form Search Engine

Luciano Barbosa, Hoa Nguyen, Thanh Nguyen, Ramesh Pinnamaneni, Juliana Freire

University of Utah

Introduction

We present DeepPeep (*http://www.deeppeep.org*), a new search engine specialized in Web forms. DeepPeep uses a scalable infrastructure for discovering, organizing and analyzing Web forms which serve as entry points to hidden-Web sites. DeepPeep provides an intuitive interface that allows users to explore and visualize large form collections.



DeepPeep Architecture

Components

Focused Crawler To locate forms on the Web, we use the ACHE crawler [1].

- ACHE focuses its crawl on a given topic based on the contents of pages
- Prioritizes links that are more likely to lead to pages that contain searchable forms
- Automatically improves its focus strategy during a crawl by applying online learning.

Form Classifier To filter out irrelevant forms, we use HIerarchical Form Identification (HIFI) [2].

- Classifies forms with respect to a domain
- Scalable: Utilizes form features that can be automatically extracted.

Form Clustering To group together forms that belong to the same domain we use the Context-Aware Form Clustering (CAFC), a framework for clustering Web forms [3].

- Models Web forms as a set of hyperlinked objects
- Uses visible information in the form context both within and in the neighborhood of forms as the basis for similarity comparison.

Label Extraction We use LabelEx, a learning-based approach for extracting labels of form elements [4].

- Makes use of a learning classifier ensemble to identify element-label mappings
- Applies a reconciliation step which leverages the classifier-derived mappings to boost extraction accuracy.

PruSM and Schema Matching To identify the correspondences among different form attributes (and their labels) we apply PruSM, a prudent schema matching algorithm [5].

- Does not require any manual pre-processing of forms
- Effectively handles both noisy data and rare labels.

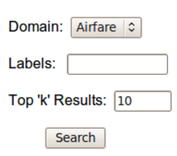
DeepPeep Site

Dataset The current version of DeepPeep contains 46,000 forms in 7 domains: Auto, Airfare, Biology, Books, Rental, Hotel and Job. We use Lucene [6] to index both the contents of the form and of the page where the form is located, as well as the form labels extracted by LabelEx.

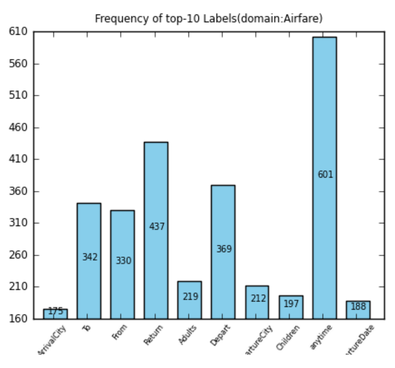
Simple Query Interface is a keyword-based interface used to retrieve forms from the DeepPeep Form Repository.

Advanced Query Interface supports *structured queries* using attribute-value pairs (e.g., *state=Utah*) and *meta-data queries* (e.g., retrieve all forms with a label "*state*").

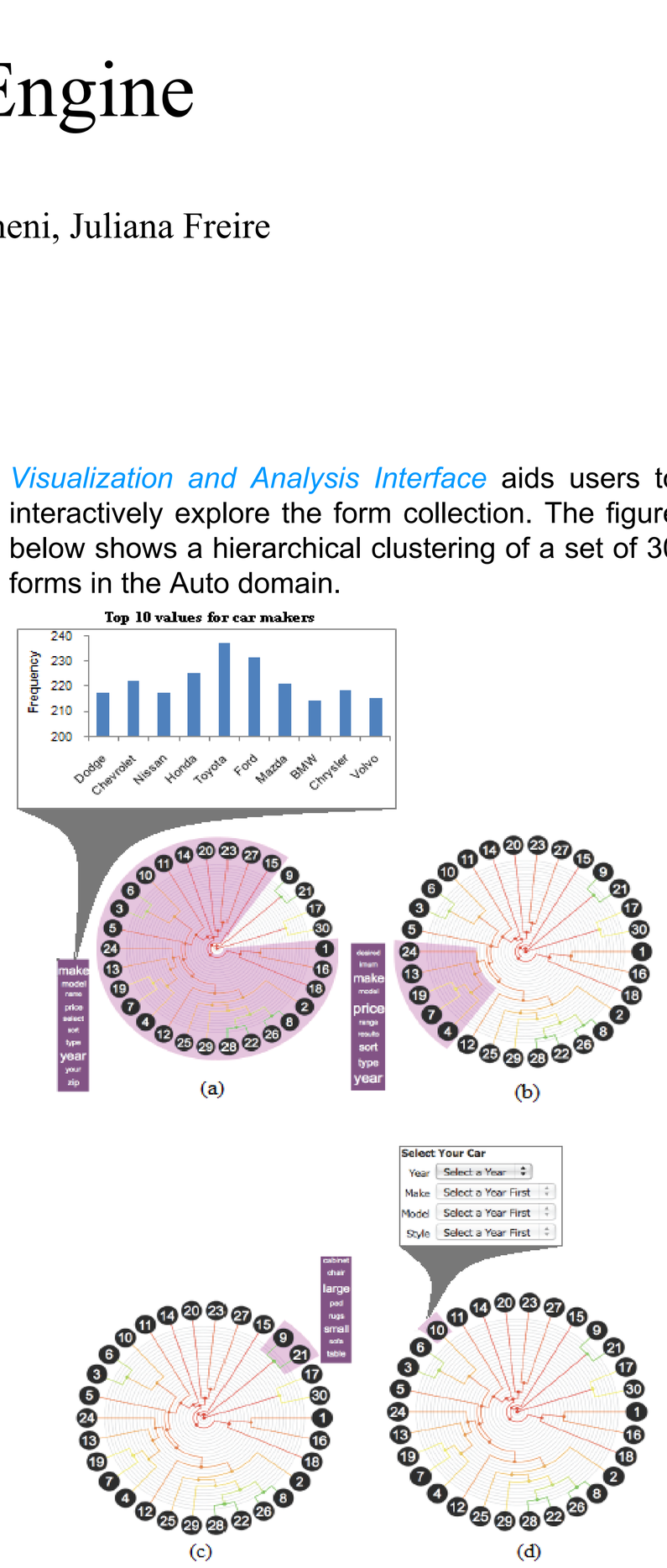
Expert Query Interface allows expert users to visualize underlying data distribution and help them explore the form collection with complex queries (e.g., show the top-k labels in a domain, or top-k values for given attributes)



Domain	Labels	Label-Freq(%)
Airfare	anytime	601(6.72%)
Airfare	Return	437(4.88%)
Airfare	Depart	369(4.12%)
Airfare	To	342(3.82%)
Airfare	From	330(3.69%)
Airfare	Adults	219(2.45%)
Airfare	Departure City	212(2.37%)
Airfare	Children	197(2.2%)
Airfare	Departure Date	188(2.1%)
Airfare	Arrival City	175(1.96%)



Visualization and Analysis Interface aids users to interactively explore the form collection. The figure below shows a hierarchical clustering of a set of 30 forms in the Auto domain.



Conclusion

We presented the overall architecture of DeepPeep which can support both general and specific deep Web search; benefits not only casual users but also application builders. The system provides a scalable and automatic solution to deep Web search and can adapt to the dynamic evolution of deep Web which is growing fast and will play an important role in the future of search .

References

[1] L. Barbosa and J. Freire. An adaptive crawler for locating hidden-web entry points. In WWW, pages 441–450, 2007.
[2] L. Barbosa and J. Freire. Combining classifiers to identify online databases. In WWW, pages 431–440, 2007.
[3] L. Barbosa, J. Freire, and A. Silva. Organizing hidden-web databases by clustering visible web documents. In ICDE, pages 621–633, 2007.
[4] H. Nguyen, T. Nguyen, and J. Freire. Learning to extract form labels. In VLDB, pages 684–694, 2008.
[5] T. Nguyen. Prudent schema matching for web forms. Technical report, University of Utah, 2008.
[6] Lucene. lucene.apache.org.

Acknowledgments

This work has been partially supported by the National Science Foundation (under grants IIS-0713637, IIS-0746500, CNS-0751152, IIS-0534628) and a University of Utah Seed Grant.